Analysis of multinomial counts with joint zero-inflation, with an application to health economics

Alpha Oumar Diallo, Aliou Diop, Jean-François Dupuy

Analysis of multinomial counts with joint zero-inflation, with an application to health economics

Alpha Oumar DIALLO[a,b], Aliou DIOP[a], Jean-François DUPUY[b]

[a] *LERSTAD, CEA-MITIC, Gaston Berger University, Saint Louis, Senegal.*
[b] *IRMAR-INSA, Rennes, France.*

Abstract

Zero-inflated regression models for count data are often used in health economics to analyse demand for medical care. Indeed, excess of zeros often affects health-care utilization data. Much of the recent econometric literature on the topic has focused on univariate health-care utilization measures, such as the number of doctor visits. However, health service utilization is usually measured by a number of different counts (*e.g.*, numbers of visits to different health-care providers). In this case, zero-inflation may jointly affect several of the utilization measures. In this paper, a zero-inflated regression model for multinomial counts with joint zero-inflation is proposed. Maximum likelihood estimators in this model are constructed and their properties are investigated, both theoretically and numerically. We apply the proposed model to an analysis of health-care utilization.

Keywords: excess zeros, health-care utilization, multinomial logit.

1. Introduction

Statistical modeling of count data with zero inflation has become an important issue in numerous fields and in particular, in econometrics. The zero inflation (or excess zeros) problem occurs when the proportion of zero counts in the observed sample is much larger than predicted by standard count models. In health economics, this issue often arises in analysis of health-care utilization, as measured by the number of doctor visits (Sarma and Simpson, 2006; Sari, 2009; Staub and Winkelmann, 2013). The present work is also motivated by an econometric analysis of health-care utilization and is illustrated by a data set described by Deb and Trivedi (1997).

Deb and Trivedi (1997) investigate the demand for medical care by elderlies in the United States. Their analysis is based on data from the National Medical Expenditure Survey (NMES) conducted in 1987 and 1988. These data provide a comprehensive picture of how Americans (aged 66 years and over) use and pay for health services. Six measures of health-care utilization were reported in this study, namely the number of visits to a doctor in an office setting, the number of visits to a non-doctor health professional (such as a nurse, optician, physiotherapist...) in an office setting, the number of visits to a doctor in an outpatient setting, the number of visits to a non-doctor in an outpatient setting, the number of visits to an emergency service and the number of hospital stays. A feature of these data is the high proportion of zero counts observed for some of the health-care utilization measures, that is, there is a high proportion of non-users of the corresponding health-care service over the study period. In addition to health services utilization, the data set also provides information on health status, sociodemographic characteristics and economic status. Deb and Trivedi (1997) analyse separately each measure of health-care utilization by fitting models for zero-inflated count data to each type of health-care usage in turns. However, several studies suggest that health-care utilization measures are not independent (Gurmu and Elder, 2000; Wang, 2003). Therefore,

Email addresses: Alpha-Oumar.Diallo1@insa-rennes.fr (Alpha Oumar DIALLO), aliou.diop@ugb.edu.sn (Aliou DIOP), Jean-Francois.Dupuy@insa-rennes.fr (Jean-François DUPUY)

we suggest to analyse jointly the various health-care utilization measures by fitting a multinomial logistic regression model to the data.

For illustrative purpose, and in order to keep notations simple, we will illustrate our model and methodology by considering three out of the six measures of health-care utilization, namely the: i) number Z_1 of consultations with a non-doctor in an office setting (denoted by *ofnd* in what follows), ii) number Z_2 of consultations with a non-doctor in an outpatient setting (*opnd*) and iii) number Z_3 of consultations with a doctor in an office setting (*ofd*). If m_i denotes the total number of consultations for the i-th individual and $\mathbf{X}_i$ is a vector of covariates for this individual, we let $Z_i = (Z_{1i}, Z_{2i}, Z_{3i})$ and we assume that Z_i has a multinomial distribution $\text{mult}(m_i, \mathbf{p}_i)$, where $\mathbf{p}_i = (p_{1i}, p_{2i}, p_{3i})$, $p_{1i} = \mathbb{P}(Z_{1i} = 1|\mathbf{X}_i)$ is the probability that a consultation is of type *ofnd*, $p_{2i} = \mathbb{P}(Z_{2i} = 1|\mathbf{X}_i)$ is the probability that a consultation is of type *opnd* and $p_{3i} = \mathbb{P}(Z_{3i} = 1|\mathbf{X}_i)$ is the probability that a consultation is of type *ofd*. We consider individuals in the NMES data set who have a total number of consultations less than or equal to 25. Among these 3224 individuals, frequencies of zero in variables *ofnd*, *opnd* and *ofd* are 62.7%, 81.3% and 1.5% respectively. Frequencies of zeros occuring simultaneously in variables of pairs (*ofnd* and *opnd*), (*ofnd* and *ofd*) and (*opnd* and *ofd*) are 51.7%, 0.24% and 1% respectively. That is, 51.7% of the surveyed subjects did not use any services associated with counts Z_1 and Z_2. This high frequency and the very low frequency of zero counts for *ofd* suggest that there may exist some permanent non-users of *ofnd* and *opnd*, *i.e.*, individuals who would never use these health-care services. In other words, there may exist an excess of observations of the form $(0, 0, m_i)$ in the data set.

To accommodate these observations, we propose to define, for each individual i, a zero-inflated multinomial regression model as the mixture

$$\pi_i \cdot \delta_{(0,0,m_i)} + (1 - \pi_i) \cdot \text{mult}(m_i, \mathbf{p}_i) \tag{1.1}$$

of the multinomial distribution $\text{mult}(m_i, \mathbf{p}_i)$ with a degenerate distribution $\delta_{(0,0,m_i)}$ at $(0, 0, m_i)$. π_i represents the probability that the i-th individual is a permanent non-user of health-care services of the type *ofnd* and *opnd*.

Mixture models for zero-inflated count data date back to early '90s. Zero-inflated Poisson (ZIP) regression was proposed by Lambert (1992) and further developed by Dietz and Böhning (2000), Li (2011), Lim *et al.* (2014) and Monod (2014), among many others. Zero-inflated negative binomial (ZINB) regression was proposed by Ridout *et al.* (2001), see also Moghimbeigi *et al.* (2008), Mwalili *et al.* (2008), Garay *et al.* (2011). Hall (2000) and Vieira *et al.* (2000) introduced the zero-inflated binomial (ZIB) regression model, see also Diop *et al.* (2016). But to the best of our knowledge, and although some related models can be found in Kelley and Anderson (2008) and Bagozzi (2015), the zero-inflated multinomial model (1.1) has not been yet considered. Kelley and Anderson (2008) (respectively Bagozzi, 2015) propose a model for a discrete ordinal (respectively nominal) dependent variable with levels $\{0, 1, \ldots, J\}$ and zero-inflation. However, authors do not report any systematic investigation of their models (such as model identifiability or estimation). In the present paper, we aim at providing a rigorous study of model (1.1) that will serve as a basis for future application of the model to real-data problems. We derive maximum likelihood estimators of parameters π_i and $\mathbf{p}_i$, we establish their asymptotic properties (consistency and asymptotic normality) and we assess their finite-sample behaviour using simulations. Then, we illustrate the model on the health-care utilization data set described above.

The remainder of the paper is organized as follows. In Section 2, we specify precisely the model and we address the estimation of π_i and $\mathbf{p}_i$. In Section 3, we report results of our simulation study. Section 4 describes the health-care data analysis. A conclusion and some perspectives are provided in Section 5. All proofs are postponed to an appendix.

2. Zero-inflated multinomial regression model

In this section, we describe the zero-inflated multinomial (ZIM) regression model. We consider two cases: i) π_i is fixed (that is, $\pi_i = \pi$ for every individual) and ii) π_i depends on covariates. In section 2.3, identifiability of the ZIM model and asymptotics of the maximum likelihood estimator are described for fixed

π but results can be generalized to case ii) without major difficulty. Moreover, for notational simplicity, we consider the case where the multinomial response Z_i has $K = 3$ outcomes. Proofs for a general K proceed similarly.

2.1. Model and estimation with fixed π

Let $(Z_i, \mathbf{X}_i)$, $i = 1, \ldots, n$ be independent random vectors defined on the probability space $(\Omega, \mathcal{C}, \mathbb{P})$. For every i, we assume that given the total $Z_{1i} + Z_{2i} + Z_{3i} = m_i$, the multivariate response $Z_i = (Z_{1i}, Z_{2i}, Z_{3i})$ is generated from the model

$$Z_i \sim \begin{cases} (0, 0, m_i) & \text{with probability } \pi, \\ \text{mult}(m_i, \mathbf{p}_i) & \text{with probability } 1 - \pi, \end{cases} \tag{2.2}$$

where $\mathbf{p}_i = (p_{1i}, p_{2i}, p_{3i})$ and $p_{1i} + p_{2i} + p_{3i}$=1. This model reduces to the standard multinomial distribution (with three modalities, here) if $\pi = 0$, while $\pi > 0$ leads to simultaneous zero-inflation in the first two modalities. We model probabilities p_{1i}, p_{2i} and p_{3i} $(i = 1, \ldots, n)$ via multinomial logistic regression:

$$p_{1i} = \frac{e^{\beta_1^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}}, \quad p_{2i} = \frac{e^{\beta_2^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}} \quad \text{and} \quad p_{3i} = \frac{1}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}}, \tag{2.3}$$

where $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})^\top$ is a vector of predictors or covariates (both categorical and continuous covariates are allowed) and $\top$ denotes the transpose operator. Let $\psi = (\pi, \beta_1^\top, \beta_2^\top)^\top$ be the unknown k-dimensional parameter of ZIM model ($k := 1+2p$). For $i = 1, \ldots, n$, let $J_i := 1_{\{Z_i \neq (0,0,m_i)\}}$ and $h_i(\beta) = 1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}$, where $\beta = (\beta_1^\top, \beta_2^\top)^\top$. Then, the log-likelihood of ψ based on observations $(Z_1, \mathbf{X}_1), \ldots, (Z_n, \mathbf{X}_n)$ is:

$$\begin{aligned} l_n(\psi) &= \sum_{i=1}^{n} \left\{ (1 - J_i) \log \left(\pi + (1 - \pi) \frac{1}{(h_i(\beta))^{m_i}} \right) \right. \\ &\qquad \left. + J_i \left[\log \left(\frac{m_i!}{Z_{1i}! Z_{2i}! Z_{3i}!} \right) - m_i \log h_i(\beta) + Z_{1i} \beta_1^\top \mathbf{X}_i + Z_{2i} \beta_2^\top \mathbf{X}_i + \log(1 - \pi) \right] \right\}, \\ &:= \sum_{i=1}^{n} l_{[i]}(\psi). \end{aligned} \tag{2.4}$$

The maximum likelihood estimator $\hat{\psi}_n := (\hat{\pi}, \hat{\beta}_1^\top, \hat{\beta}_2^\top)^\top$ of ψ is the solution of the k-dimensional score equation

$$\dot{l}_n(\psi) := \frac{\partial l_n(\psi)}{\partial \psi} = 0. \tag{2.5}$$

Solving this (non-linear) equation is relatively straightforward using standard mathematical softwares. In our simulation study and real-data analysis, we use `R` package `maxLik` (Henningsen and Toomet, 2011), which provides efficient computational tools for solving likelihood equations such as (2.5).

We need to introduce some further notations and a few regularity assumptions before stating asymptotic properties of $\hat{\psi}_n$.

2.2. Some further notations

It will be useful to define the $(p \times n)$ and $(k \times 3n)$ matrices:

$$\mathbb{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{np} \end{pmatrix} \text{ and } \mathbb{V} = \begin{pmatrix} \mathbb{1}_{(1,n)} & 0_{(1,n)} & 0_{(1,n)} \\ 0_{(p,n)} & \mathbb{X} & 0_{(p,n)} \\ 0_{(p,n)} & 0_{(p,n)} & \mathbb{X} \end{pmatrix},$$

where $\mathbb{1}_{(1,n)}$ denotes the n-dimensional raw vector $(1,1,\cdots,1)$ and $0_{(a,b)}$ denotes the $(a\times b)$ matrix whose components are all equal to zero (with a and b two positive integers). Let also $C(\psi)=(C_j(\psi))_{1\leq j\leq 3n}$ be the $3n$-dimensional column vector defined by

$$C(\psi)=(A_1(\psi),\ldots,A_n(\psi),B_{1,1}(\psi),\ldots,B_{n,1}(\psi),B_{1,2}(\psi),\ldots,B_{n,2}(\psi))^\top,$$

where for every $i=1,\ldots,n$,

$$\begin{aligned}
A_i(\psi)&=\frac{(h_i(\beta))^{m_i}-1}{\pi\left[(h_i(\beta))^{m_i}-1\right]+1}(1-J_i)-\frac{1}{1-\pi}J_i,\\
B_{i,\ell}(\psi)&=-(1-\pi)\frac{m_i e^{\beta_\ell^\top\mathbf{x}_i}}{k_i(\psi)}(1-J_i)+\left(-\frac{m_i e^{\beta_\ell^\top\mathbf{x}_i}}{h_i(\beta)}+Z_{\ell i}\right)J_i,\quad \ell=1,2,
\end{aligned}$$

and $k_i(\psi)=\pi\left[(h_i(\beta))^{m_i+1}-h_i(\beta)\right]+h_i(\beta)$. Then, some simple algebra shows that the likelihood equation (2.5) can be rewritten as

$$\dot{l}_n(\psi)=\mathbb{V}C(\psi)=0.$$

If $A=(A_{ij})_{1\leq i\leq a,1\leq j\leq b}$ is some $(a\times b)$ matrix, we denote its j-th column ($j=1,\ldots,b$) by $A_{\bullet j}$. That is, $A_{\bullet j}=(A_{1j},\ldots,A_{aj})^\top$. Then, it will be useful to rewrite the score vector as

$$\dot{l}_n(\psi)=\sum_{j=1}^{3n}\mathbb{V}_{\bullet j}C_j(\psi).$$

We shall further denote by $\ddot{l}_n(\psi)=\partial^2 l_n(\psi)/\partial\psi\partial\psi^\top$ the $(k\times k)$ matrix of second derivatives of $l_n(\psi)$. Let $\mathbb{D}(\psi)=(\mathbb{D}_{ij}(\psi))_{1\leq i,j\leq 3n}$ be the $(3n\times 3n)$ block matrix defined as

$$\mathbb{D}(\psi)=\begin{pmatrix}\mathbb{D}_1(\psi) & \mathbb{D}_4(\psi) & \mathbb{D}_5(\psi)\\ \mathbb{D}_4(\psi) & \mathbb{D}_2(\psi) & \mathbb{D}_6(\psi)\\ \mathbb{D}_5(\psi) & \mathbb{D}_6(\psi) & \mathbb{D}_3(\psi)\end{pmatrix},$$

where $\mathbb{D}_1(\psi)$ to $\mathbb{D}_6(\psi)$ are $(n\times n)$ diagonal matrices, with i-th diagonal elements respectively given by

$$\begin{aligned}
\mathbb{D}_{1,ii}(\psi)&=\left(\frac{(h_i(\beta))^{m_i}-1}{\pi\left[(h_i(\beta))^{m_i}-1\right]+1}\right)^2(1-J_i)+\frac{1}{(1-\pi)^2}J_i,\\
\mathbb{D}_{\ell+1,ii}(\psi)&=\frac{(1-\pi)(1-J_i)e^{\beta_\ell^\top\mathbf{x}_i}\left((k_i(\psi)-e^{\beta_\ell^\top\mathbf{x}_i}(\pi[(m_i+1)(h_i(\beta))^{m_i}-1]+1)\right)}{(k_i(\psi))^2}\\
&\qquad-\frac{m_i J_i e^{\beta_\ell^\top\mathbf{x}_i}\left(h_i(\beta)-e^{\beta_\ell^\top\mathbf{x}_i}\right)}{(h_i(\beta))^2},\quad \ell=1,2,\\
\mathbb{D}_{\ell+3,ii}(\psi)&=-\frac{(1-J_i)m_i e^{\beta_\ell^\top\mathbf{x}_i}(h_i(\beta))^{m_i+1}}{(k_i(\psi))^2},\quad \ell=1,2,\\
\mathbb{D}_{6,ii}(\psi)&=-\frac{(1-\pi)(1-J_i)m_i e^{\beta_1^\top\mathbf{x}_i}e^{\beta_2^\top\mathbf{x}_i}\left(\pi[(h_i(\beta))^{m_i}-1]+1\right)}{(k_i(\psi))^2}-\frac{J_i m_i e^{\beta_1^\top\mathbf{x}_i}e^{\beta_2^\top\mathbf{x}_i}}{(h_i(\beta))^2}.
\end{aligned}$$

Then, some tedious albeit not difficult algebra shows that $\ddot{l}_n(\psi)$ can be expressed as $\ddot{l}_n(\psi)=-\mathbb{V}\mathbb{D}(\psi)\mathbb{V}^\top$. Note that $C(\psi),\mathbb{V}$ and $\mathbb{D}(\psi)$ depend on n. However, in order to simplify notations, n will not be used as a lower index for these quantities. In the next section, we state some regularity conditions and asymptotic properties of the maximum likelihood estimator $\hat{\psi}_n$.

2.3. Regularity conditions, model identifiability and asymptotic results

The following conditions are somewhat classical in the framework of generalized linear regression models and are adapted to our setting.

C1 Covariates X_{ij} are bounded and $\text{var}[X_{ij}] > 0$, for every $i = 1, 2, \ldots$ and $j = 2, \ldots, p$. The X_{ij} $(j = 1, \ldots, p)$ are linearly independent, for every $i = 1, 2, \ldots$

C2 The true parameter value $\psi_0 := (\pi_0, \beta_{1,0}^\top, \beta_{2,0}^\top)^\top$ lies in the interior of some known compact set $\mathbf{K} \subset [0,1] \times \mathbb{R}^p \times \mathbb{R}^p$ (in what follows, we will also note $\beta_0 := (\beta_{1,0}^\top, \beta_{2,0}^\top)^\top$).

C3 The Hessian matrix $\ddot{l}_n(\psi)$ is negative definite and of full rank, for every $n = 1, 2, \ldots$ and $\frac{1}{n}\ddot{l}_n(\psi)$ converges to a negative definite matrix. Let λ_n and Λ_n be respectively the smallest and largest eigenvalues of $\mathbb{VD}(\psi_0)\mathbb{V}^\top$. There exists a finite positive constant c_1 such that $\Lambda_n/\lambda_n < c_1$ for every $n = 1, 2, \ldots$ The matrix $\mathbb{V}\mathbb{V}^\top$ is positive definite for every $n = 1, 2, \ldots$ and its smallest eigenvalue $\tilde{\lambda}_n$ tends to $+\infty$ as $n \to \infty$.

Next condition will be useful for proving identifiability of the ZIM model (*i.e.*, distinct parameter values yield distinct values of the likelihood function).

C4 For every $i = 1, \ldots, n$, we have $m_i \geq 2$ (that is, in our application, we consider individuals who had at least two visits of all type).

Then, the following result holds for a fixed probability of zero-inflation (proof is given in Appendix A):

Theorem 2.1 (Identifiability). *Under conditions C1-C4, the ZIM model* (2.2)-(2.3) *is identifiable, that is,* $l_{[i]}(\psi) = l_{[i]}(\psi^*)$ *almost surely implies* $\psi = \psi^*$.

Now, we state asymptotic properties of the estimator $\widehat{\psi}_n$. Proofs are outlined in Appendix B. Main steps are similar to proofs of asymptotics in the logistic regression model (*e.g.*, Gouriéroux and Monfort, 1981). However, specific technical difficulties arise in the ZIM model. In particular, observations $(Z_i, \mathbf{X}_i)$ are not identically distributed (the number m_i of visits of all types varies across individuals).

In what follows, the space $\mathbb{R}^k$ of k-dimensional vectors is equipped with the Euclidean norm $\|\cdot\|$. The space of $(k \times k)$ real matrices is equipped with the norm $\|A\|_2 := \max_{\|x\|=1} \|Ax\|$ (for notations simplicity, we use $\|\cdot\|$ for both norms). Recall that for a symmetric real $(k \times k)$-matrix A with eigenvalues $\lambda_1, \ldots, \lambda_k$, $\|A\| := \|A\|_2 = \max_i |\lambda_i|$. Finally, we let I_k denote the identity matrix of order k. Our results are as follows:

Theorem 2.2 (Existence and consistency). *The maximum likelihood estimator* $\widehat{\psi}_n$ *exists almost surely as* $n \to \infty$ *and converges almost surely to* ψ_0.

Moreover, $\widehat{\psi}_n$ is asymptotically Gaussian:

Theorem 2.3 (Asymptotic normality). *Let* $\widehat{\Sigma}_n := \mathbb{VD}(\widehat{\psi}_n)\mathbb{V}^\top$. *Then, as* $n \to \infty$, $\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0)$ *converges in distribution to the Gaussian vector* $\mathcal{N}(0, I_k)$.

In the next section, we describe briefly the ZIM model with covariate-dependent probability of zero-inflation.

2.4. Model and estimation with covariate-dependent π_i

We assume that the probability π_i of $(0, 0, m_i)$-inflation for individual i depends on some observed q-dimensional covariate $\mathbf{W}_i$ ($\mathbf{W}_i$ may overlap with $\mathbf{X}_i$ or be distinct from $\mathbf{X}_i$. This issue is discussed in the application). We model π_i via logistic regression:

$$\pi_i = \frac{e^{\gamma^\top \mathbf{W}_i}}{1 + e^{\gamma^\top \mathbf{W}_i}}. \tag{2.6}$$

The log-likelihood of $\psi = (\gamma^\top, \beta_1^\top, \beta_2^\top)^\top$, based on observations $(Z_i, \mathbf{X}_i, \mathbf{W}_i)$, $i = 1, \ldots, n$ is similar to (2.4), with π_i replacing π, and the maximum likelihood estimator of ψ is defined similarly as above. Identifiability of the ZIM model (2.2)-(2.3)-(2.6) can be proved along the same lines as Theorem 2.1 under the following additional regularity condition: covariates W_{ij} are bounded and $\text{var}[W_{ij}] > 0$, for every $i = 1, 2, \ldots$ and $j = 2, \ldots, q$. The W_{ij} $(j = 1, \ldots, q)$ are linearly independent, for every $i = 1, 2, \ldots$

3. A simulation study

In this section, we assess finite-sample properties of the maximum likelihood estimator $\hat{\psi}_n$ with fixed π and covariate-dependent π_i.

Case (i): fixed probability of zero-inflation. We simulate data from a ZIM model defined by:

$$p_{1i} = \frac{e^{\beta_1^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}}, \quad p_{2i} = \frac{e^{\beta_2^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}} \quad \text{and} \quad p_{3i} = 1 - p_{1i} - p_{2i},$$

where $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{i7})^\top$ and $X_{i2}, \ldots, X_{i7}$ are independent covariates simulated from normal $\mathcal{N}(0,1)$, uniform $\mathcal{U}(2,5)$, normal $\mathcal{N}(1,1.5)$, exponential $\mathcal{E}(1)$, binomial $\mathcal{B}(1,0.3)$ and normal $\mathcal{N}(-1,1)$ distributions respectively. Parameters β_1 and β_2 are chosen as $\beta_1 = (0.3, 1.2, 0.5, -0.75, -1, 0.8, 0)^\top$ and $\beta_2 = (0.5, 0.5, 0, -0.5, 0.5, -1.1, 0)^\top$. Several sample sizes n are considered: $n = 150, 300$ and 500. Numbers m_i are allowed to vary across subjects, with $m_i \in \{3, 4, 5\}$. Let $(n_3, n_4, n_5) = (\text{card}\{i : m_i = 3\}, \text{card}\{i : m_i = 4\}, \text{card}\{i : m_i = 5\})$. For $n = 150$, we let $(n_3, n_4, n_5)) = (50, 50, 50)$. For $n = 300$, we let $(n_3, n_4, n_5) = (120, 100, 80)$ and for $n = 500$, we let $(n_3, n_4, n_5) = (230, 170, 100)$. Zero-inflation is simulated from a Bernoulli variable with parameter π, with $\pi = 0.15, 0.25$ and 0.5.

Case (ii): covariate-dependent probability of zero-inflation. In a second set of simulation scenarios, zero-inflation is allowed to depend on covariates. Simulation design is essentially similar as above, except that: i) $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{i5})^\top$ and $X_{i2}, \ldots, X_{i5}$ are simulated independently from uniform $\mathcal{U}(2,5)$, normal $\mathcal{N}(1,1.5)$, exponential $\mathcal{E}(1)$ and binomial $\mathcal{B}(1,0.3)$ distributions respectively and ii) for each individual i, zero-inflation is simulated from a Bernoulli random variable with parameter π_i, where $\text{logit}(\pi_i) = \gamma^\top \mathbf{W}_i$ and $\mathbf{W}_i$ is taken as $\mathbf{W}_i = (1, X_{i2}, X_{i5}, W_{i4})^\top$ with $W_{i4} \sim \mathcal{N}(-1,1)$. Parameters β_1 and β_2 are taken as $\beta_1 = (0.3, 0.5, -0.75, -1, 0)^\top$ and $\beta_2 = (0.5, 0, -0.5, 1.5, -1.1)^\top$. The parameter vector $\gamma \in \mathbb{R}^4$ is chosen to yield various average proportions of zero-inflation within each sample, namely: $0.15, 0.25$ and 0.5.

Results. For each combination `sample size` $\times$ `zero-inflation proportion`, we simulate $N = 5000$ samples and for each of them, we calculate the maximum likelihood estimate $\hat{\psi}_n$ of (π, β_1, β_2) (case (i)) and $(\gamma, \beta_1, \beta_2)$ (case (ii)). Several authors developed EM-type algorithms for estimation in zero-inflated models (*e.g.*, Wang, 2003; Kelley and Anderson, 2008). Other authors proceed to direct maximization using Newton-Raphson or related algorithms (*e.g.*, Staub and Winkelmann, 2013). Here, we use Newton-Raphson-like algorithm implemented in the `R` package `maxLik` developed by Henningsen and Toomet (2011).

Based on the N estimates, we obtain, for each simulation scenario, the: i) empirical biais of each estimator, ii) average standard error (SE) and empirical standard deviation (SD) of each estimator, iii) empirical coverage probability (CP) and average length ℓ(CI) of 95%-level confidence interval for each parameter. For case (i), results are given in Table 1 ($\pi = 0.15$), Table 2 ($\pi = 0.25$) and Table 3 ($\pi = 0.5$). For case (ii), results are given in Table 4 (average sample proportion of zero-inflation equal to 0.15), Table 5 (average sample proportion of zero-inflation equal to 0.25) and Table 6 (average sample proportion of zero-inflation equal to 0.5).

From these tables, the bias, SE, SD and ℓ(CI) of all estimators decrease as sample size increases. The bias stays moderate and empirical coverage probabilities are close to the nominal confidence level. Maximum likelihood seems to provide an efficient method for estimating ZIM model, even when the number of parameters is quite large.

Finally, we assess empirically the Gaussian approximation stated in Theorem 2.3 by plotting normal Q-Q plots of the estimates. Figures 1, 2 and 3 provide plots for case (ii) with $n = 300$ and an average sample proportion of zero-inflation equal to 0.25 (plots for the other simulation scenarios are similar and are thus omitted). From these plots, the distribution of the maximum likelihood estimator in the ZIM model is reasonably approximated by the Gaussian distribution. This, in particular, will allow Wald tests of covariate effects to be performed in ZIM model.

n		$\widehat{\pi}$	$\widehat{\beta}_1$							$\widehat{\beta}_2$						
			$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{1,6}$	$\widehat{\beta}_{1,7}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$	$\widehat{\beta}_{2,6}$	$\widehat{\beta}_{2,7}$
150																
	bias	-0.0020	-0.0073	0.0389	0.0157	-0.0225	-0.0282	0.0259	-0.0028	-0.0072	0.0224	0.0036	-0.0149	0.0158	-0.0284	-0.0035
	SD	0.0320	0.6687	0.1728	0.1749	0.1171	0.2110	0.3208	0.1475	0.6542	0.1630	0.1735	0.1131	0.1577	0.3751	0.1469
	SE	0.0316	0.6470	0.1689	0.1689	0.1109	0.2046	0.3153	0.1446	0.6434	0.1614	0.1699	0.1085	0.1531	0.3716	0.1465
	CP	0.9300	0.9462	0.9478	0.9424	0.9404	0.9460	0.9498	0.9466	0.9436	0.9500	0.9486	0.9430	0.9474	0.9482	0.9500
	ℓ(CI)	0.1234	2.5286	0.6597	0.6604	0.4330	0.7988	1.2326	0.5644	2.5137	0.6299	0.6643	0.4236	0.5939	1.4493	0.5718
300																
	bias	-0.0011	0.0006	0.0182	0.0075	-0.0113	-0.0142	0.0135	0.0002	-0.0033	0.0102	0.0030	-0.0083	0.0080	-0.0131	0.0007
	SD	0.0224	0.4495	0.1210	0.1185	0.0782	0.1445	0.2180	0.1014	0.4473	0.1141	0.1192	0.0768	0.1061	0.2590	0.1025
	SE	0.0224	0.4527	0.1178	0.1181	0.0774	0.1422	0.2204	0.1007	0.4497	0.1123	0.1187	0.0758	0.1050	0.2583	0.1020
	CP	0.9408	0.9538	0.9432	0.9496	0.9476	0.9428	0.9506	0.9508	0.9560	0.9476	0.9458	0.9470	0.9476	0.9498	0.9478
	ℓ(CI)	0.0878	1.7719	0.4610	0.4624	0.3028	0.5563	0.8631	0.3939	1.7599	0.4393	0.4648	0.2965	0.4095	1.0105	0.3990
500																
	bias	-0.0009	0.0006	0.0115	0.0041	-0.0068	-0.0073	0.0065	0.0017	-0.0019	0.0064	0.0010	-0.0054	0.0049	-0.0093	0.0001
	SD	0.0175	0.3519	0.0929	0.0914	0.0604	0.1115	0.1739	0.0804	0.3506	0.0884	0.0922	0.0579	0.0830	0.2065	0.0799
	SE	0.0175	0.3534	0.0918	0.0922	0.0603	0.1107	0.1719	0.0784	0.3512	0.0875	0.0927	0.0590	0.0814	0.2012	0.0794
	CP	0.9456	0.9524	0.9462	0.9502	0.9506	0.9514	0.9434	0.9442	0.9496	0.9510	0.9516	0.9522	0.9452	0.9440	0.9516
	ℓ(CI)	0.0685	1.3841	0.3594	0.3611	0.2360	0.4336	0.6732	0.3069	1.3752	0.3426	0.3632	0.2310	0.3183	0.7878	0.3109

Table 1: Simulation results (case (i), $\pi = 0.15$). SE: average standard error. SD: empirical standard deviation. CP: empirical coverage probability of 95%-level confidence intervals. ℓ(CI): average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

n		$\widehat{\pi}$	$\widehat{\beta}_1$							$\widehat{\beta}_2$						
			$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{1,6}$	$\widehat{\beta}_{1,7}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$	$\widehat{\beta}_{2,6}$	$\widehat{\beta}_{2,7}$
150																
	bias	-0.0044	-0.0014	0.0418	0.0180	-0.0314	-0.0297	0.0152	0.0013	0.0113	0.0220	0.0023	-0.0247	0.0185	-0.0554	0.0013
	SD	0.0379	0.7106	0.1877	0.1874	0.1243	0.2278	0.3487	0.1614	0.7164	0.1815	0.1884	0.1228	0.1718	0.4151	0.1659
	SE	0.0377	0.7003	0.1826	0.1828	0.1205	0.2201	0.3403	0.1562	0.6972	0.1751	0.1842	0.1183	0.1655	0.4023	0.1588
	CP	0.9392	0.9492	0.9454	0.9432	0.9408	0.9460	0.9476	0.9446	0.9460	0.9474	0.9448	0.9418	0.9438	0.9454	0.9466
	ℓ(CI)	0.1474	2.7343	0.7127	0.7142	0.4704	0.8584	1.3294	0.6092	2.7216	0.6828	0.7196	0.4616	0.6414	1.5672	0.6191
300																
	bias	-0.0024	-0.0061	0.0206	0.0095	-0.0157	-0.0146	0.0082	-0.0020	0.0014	0.0114	0.0003	-0.0115	0.0106	-0.0238	-0.0032
	SD	0.0267	0.4926	0.1299	0.1270	0.0855	0.1545	0.2414	0.1103	0.4930	0.1235	0.1303	0.0833	0.1133	0.2795	0.1123
	SE	0.0267	0.4865	0.1269	0.1271	0.0836	0.1525	0.2366	0.1083	0.4844	0.1214	0.1281	0.0821	0.1133	0.2778	0.1099
	CP	0.9454	0.9456	0.9420	0.9502	0.9474	0.9480	0.9466	0.9488	0.9464	0.9438	0.9444	0.9452	0.9508	0.9498	0.9478
	ℓ(CI)	0.1048	1.9036	0.4965	0.4974	0.3272	0.5965	0.9263	0.4234	1.8950	0.4748	0.5014	0.3210	0.4418	1.0861	0.4297
500																
	bias	-0.0009	0.0044	0.0146	0.0042	-0.0093	-0.0071	0.0060	0.0006	0.0036	0.0077	-0.0005	-0.0064	0.0081	-0.0167	0.0013
	SD	0.0210	0.3847	0.0984	0.1010	0.0654	0.1162	0.1847	0.0853	0.3771	0.0942	0.1002	0.0646	0.0880	0.2166	0.0868
	SE	0.0208	0.3797	0.0988	0.0991	0.0651	0.1185	0.1848	0.0843	0.3781	0.0945	0.0999	0.0639	0.0876	0.2167	0.0856
	CP	0.9488	0.9480	0.9508	0.9442	0.9460	0.9538	0.9504	0.9490	0.9506	0.9494	0.9474	0.9510	0.9496	0.9458	0.9492
	ℓ(CI)	0.0816	1.4867	0.3869	0.3882	0.2548	0.4640	0.7238	0.3300	1.4802	0.3699	0.3912	0.2500	0.3422	0.8482	0.3349

Table 2: Simulation results (case (i), $\pi = 0.25$). SE: average standard error. SD: empirical standard deviation. CP: empirical coverage probability of 95%-level confidence intervals. ℓ(CI): average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

n		$\widehat{\pi}$	$\widehat{\beta}_1$							$\widehat{\beta}_2$						
			$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{1,6}$	$\widehat{\beta}_{1,7}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$	$\widehat{\beta}_{2,6}$	$\widehat{\beta}_{2,7}$
150																
	bias	-0.0044	0.0018	0.0748	0.0270	-0.0501	-0.0479	0.0379	-0.0040	0.0056	0.0453	0.0053	-0.0360	0.0282	-0.0723	-0.0018
	SD	0.0430	0.9224	0.2484	0.2445	0.1631	0.2952	0.4615	0.2110	0.9107	0.2426	0.2454	0.1620	0.2332	0.5426	0.2144
	SE	0.0430	0.9060	0.2377	0.2364	0.1566	0.2842	0.4410	0.2030	0.9061	0.2295	0.2390	0.1543	0.2182	0.5262	0.2070
	CP	0.9462	0.9454	0.9408	0.9446	0.9424	0.9495	0.9426	0.9450	0.9521	0.9386	0.9468	0.9384	0.9434	0.9529	0.9460
	ℓ(CI)	0.1685	3.5248	0.9238	0.9210	0.6087	1.1039	1.7160	0.7885	3.5223	0.8908	0.9305	0.5995	0.8383	2.0361	0.8033
300																
	bias	-0.0025	-0.0022	0.0393	0.0132	-0.0209	-0.0267	0.0223	-0.0001	-0.0037	0.0211	0.0031	-0.0137	0.0144	-0.0306	0.0004
	SD	0.0304	0.6204	0.1669	0.1618	0.1089	0.1985	0.3078	0.1388	0.6099	0.1588	0.1630	0.1068	0.1474	0.3579	0.1420
	SE	0.0304	0.6140	0.1609	0.1604	0.1059	0.1922	0.2989	0.1370	0.6125	0.1546	0.1619	0.1043	0.1442	0.3523	0.1393
	CP	0.9460	0.9470	0.9416	0.9462	0.9426	0.9466	0.9450	0.9516	0.9516	0.9430	0.9480	0.9480	0.9494	0.9544	0.9470
	ℓ(CI)	0.1190	2.3985	0.6285	0.6268	0.4136	0.7502	1.1683	0.5347	2.3919	0.6035	0.6326	0.4072	0.5600	1.3742	0.5436
500																
	bias	-0.0007	-0.0069	0.0205	0.0096	-0.0147	-0.0103	0.0059	-0.0007	-0.0039	0.0134	0.0035	-0.0114	0.0117	-0.0242	-0.0001
	SD	0.0236	0.4795	0.1259	0.1253	0.0834	0.1521	0.2333	0.1083	0.4812	0.1213	0.1267	0.0822	0.1127	0.2756	0.1100
	SE	0.0235	0.4747	0.1243	0.1240	0.0819	0.1478	0.2311	0.1057	0.4736	0.1194	0.1252	0.0807	0.1101	0.2717	0.1074
	CP	0.9446	0.9496	0.9468	0.9510	0.9452	0.9416	0.9484	0.9472	0.9456	0.9454	0.9470	0.9450	0.9498	0.9478	0.9432
	ℓ(CI)	0.0923	1.8573	0.4863	0.4853	0.3205	0.5779	0.9046	0.4134	1.8525	0.4670	0.4900	0.3156	0.4295	1.0624	0.4200

Table 3: Simulation results (case (i), $\pi = 0.50$). SE: average standard error. SD: empirical standard deviation. CP: empirical coverage probability of 95%-level confidence intervals. ℓ(CI): average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

n		$\widehat{\gamma}$				$\widehat{\beta}_1$					$\widehat{\beta}_2$				
		$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_4$	$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$
150															
	bias	-0.0815	0.0160	-0.0982	0.0492	0.0161	0.0077	-0.0185	-0.0166	-0.0063	0.0073	0.0001	-0.0149	0.0447	-0.0258
	SD	1.0758	0.2874	0.6583	0.2867	0.6900	0.1858	0.1192	0.3212	0.3173	0.6565	0.1796	0.1149	0.2508	0.3285
	SE	1.1344	0.3004	0.7909	0.2774	0.6652	0.1790	0.1174	0.3177	0.3152	0.6453	0.1754	0.1119	0.2524	0.3273
	CP	0.9686	0.9665	0.9703	0.9502	0.9420	0.9430	0.9486	0.9490	0.9490	0.9480	0.9457	0.9476	0.9554	0.9490
	ℓ(CI)	4.3992	1.1665	2.5944	1.0756	2.5969	0.6992	0.4582	1.2409	1.2317	2.5200	0.6854	0.4368	0.9826	1.2784
300															
	bias	-0.0594	0.0106	-0.0540	0.0214	-0.0040	0.0062	-0.0091	-0.0026	-0.0012	0.0035	0.0000	-0.0087	0.0261	-0.0163
	SD	0.7920	0.2110	0.4482	0.1942	0.4662	0.1258	0.0827	0.2255	0.2259	0.4485	0.1230	0.0788	0.1770	0.2307
	SE	0.7785	0.2059	0.4431	0.1890	0.4662	0.1253	0.0820	0.2219	0.2208	0.4528	0.1229	0.0781	0.1760	0.2294
	CP	0.9487	0.9493	0.9594	0.9497	0.9505	0.9479	0.9505	0.9483	0.9449	0.9505	0.9515	0.9521	0.9481	0.9531
	ℓ(CI)	3.0378	0.8041	1.7149	0.7374	1.8246	0.4904	0.3206	0.8685	0.8645	1.7720	0.4812	0.3056	0.6878	0.8981
500															
	bias	-0.0405	0.0072	-0.0350	0.0077	0.0035	0.0026	-0.0063	-0.0004	-0.0013	-0.0002	0.0006	-0.0051	0.0158	-0.0082
	SD	0.6003	0.1589	0.3379	0.1477	0.3673	0.0995	0.0647	0.1735	0.1774	0.3539	0.0964	0.0617	0.1387	0.1806
	SE	0.5957	0.1576	0.3344	0.1438	0.3668	0.0986	0.0642	0.1742	0.1736	0.3564	0.0968	0.0612	0.1381	0.1802
	CP	0.9536	0.9536	0.9562	0.9466	0.9486	0.9466	0.9480	0.9518	0.9454	0.9538	0.9464	0.9508	0.9502	0.9504
	ℓ(CI)	2.3281	0.6161	1.3020	0.5622	1.4364	0.3862	0.2515	0.6824	0.6802	1.3957	0.3791	0.2396	0.5403	0.7060

Table 4: Simulation results (case (*ii*), average sample proportion of zero-inflation equal to 0.15). SE: average standard error. SD: empirical standard deviation. CP: empirical coverage probability of 95%-level confidence intervals. ℓ(CI): average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

n		$\widehat{\gamma}$				$\widehat{\beta}_1$					$\widehat{\beta}_2$				
		$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_4$	$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$
150															
	bias	-0.0135	-0.0272	0.0265	-0.0026	0.0100	0.0098	-0.0224	-0.0130	-0.0015	-0.0013	0.0022	-0.0153	0.0453	-0.0227
	SD	1.0894	0.3109	0.5223	0.2544	0.7035	0.1850	0.1222	0.3216	0.3466	0.6805	0.1807	0.1157	0.2592	0.3604
	SE	1.0610	0.3037	0.5125	0.2522	0.6884	0.1825	0.1191	0.3216	0.3355	0.6637	0.1774	0.1134	0.2566	0.3526
	CP	0.9534	0.9566	0.9604	0.9570	0.9488	0.9480	0.9426	0.9560	0.9454	0.9488	0.9474	0.9446	0.9512	0.9516
	ℓ(CI)	4.1293	1.1798	1.9964	0.9812	2.6895	0.7132	0.4650	1.2562	1.3108	2.5928	0.6932	0.4425	0.9985	1.3772
300															
	bias	0.0055	-0.0158	0.0150	0.0007	-0.0033	0.0070	-0.0099	-0.0130	-0.0001	-0.0001	0.0014	-0.0072	0.0183	-0.0137
	SD	0.7349	0.2109	0.3517	0.1751	0.4906	0.1298	0.0826	0.2282	0.2337	0.4731	0.1263	0.0800	0.1821	0.2465
	SE	0.7227	0.2064	0.3487	0.1719	0.4820	0.1277	0.0830	0.2252	0.2347	0.4647	0.1241	0.0790	0.1790	0.2467
	CP	0.9536	0.9540	0.9552	0.9518	0.9480	0.9482	0.9522	0.9460	0.9542	0.9498	0.9496	0.9482	0.9504	0.9524
	ℓ(CI)	2.8243	0.8060	1.3642	0.6718	1.8862	0.4997	0.3247	0.8813	0.9188	1.8184	0.4858	0.3089	0.6992	0.9653
500															
	bias	-0.0080	-0.0068	0.0075	-0.0018	-0.0010	0.0039	-0.0063	-0.0037	0.0014	-0.0013	0.0010	-0.0043	0.0143	-0.0083
	SD	0.5491	0.1560	0.2688	0.1346	0.3754	0.1001	0.0667	0.1770	0.1894	0.3643	0.0972	0.0629	0.1410	0.1926
	SE	0.5538	0.1578	0.2669	0.1315	0.3799	0.1006	0.0652	0.1767	0.1846	0.3665	0.0979	0.0621	0.1406	0.1940
	CP	0.9548	0.9556	0.9492	0.9478	0.9486	0.9484	0.9442	0.9526	0.9442	0.9480	0.9494	0.9466	0.9508	0.9530
	ℓ(CI)	2.1669	0.6172	1.0451	0.5145	1.4873	0.3940	0.2552	0.6919	0.7228	1.4352	0.3833	0.2430	0.5499	0.7597

Table 5: Simulation results (case (*ii*), average sample proportion of zero-inflation equal to 0.25). SE: average standard error. SD: empirical standard deviation. CP: empirical coverage probability of 95%-level confidence intervals. ℓ(CI): average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

n		$\widehat{\gamma}$				$\widehat{\beta}_1$					$\widehat{\beta}_2$				
		$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$	$\widehat{\gamma}_4$	$\widehat{\beta}_{1,1}$	$\widehat{\beta}_{1,2}$	$\widehat{\beta}_{1,3}$	$\widehat{\beta}_{1,4}$	$\widehat{\beta}_{1,5}$	$\widehat{\beta}_{2,1}$	$\widehat{\beta}_{2,2}$	$\widehat{\beta}_{2,3}$	$\widehat{\beta}_{2,4}$	$\widehat{\beta}_{2,5}$
150															
	bias	-0.0701	0.0260	-0.0455	0.0059	0.0090	0.0259	-0.0551	-0.0245	-0.0021	0.0053	0.0041	-0.0416	0.1009	-0.0467
	SD	0.7962	0.2210	0.4025	0.1898	1.0438	0.2950	0.1915	0.4979	0.4833	1.0219	0.2926	0.1868	0.4000	0.4972
	SE	0.7861	0.2167	0.3987	0.1838	0.9957	0.2818	0.1831	0.4815	0.4638	0.9731	0.2786	0.1747	0.3818	0.4721
	CP	0.9504	0.9482	0.9532	0.9498	0.9422	0.9444	0.9428	0.9492	0.9446	0.9464	0.9440	0.9390	0.9456	0.9398
	ℓ(CI)	3.0757	0.8475	1.5599	0.7189	3.8652	1.0931	0.7087	1.8688	1.8068	3.7779	1.0815	0.6752	1.4704	1.8383
300															
	bias	-0.0497	0.0175	-0.0199	0.0034	-0.0061	0.0141	-0.0242	-0.0099	-0.0054	-0.0072	0.0022	-0.0168	0.0547	-0.0211
	SD	0.5532	0.1521	0.2801	0.1280	0.6947	0.1948	0.1267	0.3296	0.3212	0.6748	0.1918	0.1221	0.2621	0.3316
	SE	0.5450	0.1501	0.2759	0.1273	0.6740	0.1905	0.1233	0.3268	0.3169	0.6588	0.1887	0.1175	0.2576	0.3224
	CP	0.9480	0.9486	0.9494	0.9526	0.9450	0.9490	0.9488	0.9542	0.9498	0.9472	0.9476	0.9458	0.9504	0.9458
	ℓ(CI)	2.1344	0.5877	1.0806	0.4984	2.6307	0.7435	0.4806	1.2756	1.2392	2.5716	0.7365	0.4579	1.0019	1.2606
500															
	bias	-0.0177	0.0061	-0.0082	0.0013	-0.0085	0.0097	-0.0134	-0.0042	-0.0024	-0.0059	0.0021	-0.0111	0.0330	-0.0113
	SD	0.4136	0.1141	0.2110	0.0990	0.5222	0.1491	0.0958	0.2543	0.2514	0.5198	0.1471	0.0931	0.2009	0.2590
	SE	0.4188	0.1152	0.2120	0.0976	0.5236	0.1477	0.0953	0.2533	0.2465	0.5128	0.1465	0.0909	0.1994	0.2510
	CP	0.9556	0.9554	0.9534	0.9500	0.9534	0.9516	0.9512	0.9486	0.9432	0.9460	0.9518	0.9482	0.9480	0.9430
	ℓ(CI)	1.6409	0.4514	0.8305	0.3824	2.0475	0.5777	0.3723	0.9906	0.9651	2.0052	0.5731	0.3550	0.7781	0.9826

Table 6: Simulation results (case (*ii*), average sample proportion of zero-inflation equal to 0.50). SE: average standard error. SD: empirical standard deviation. CP: empirical coverage probability of 95%-level confidence intervals. ℓ(CI): average length of confidence intervals. All results are based on $N = 5000$ simulated samples.

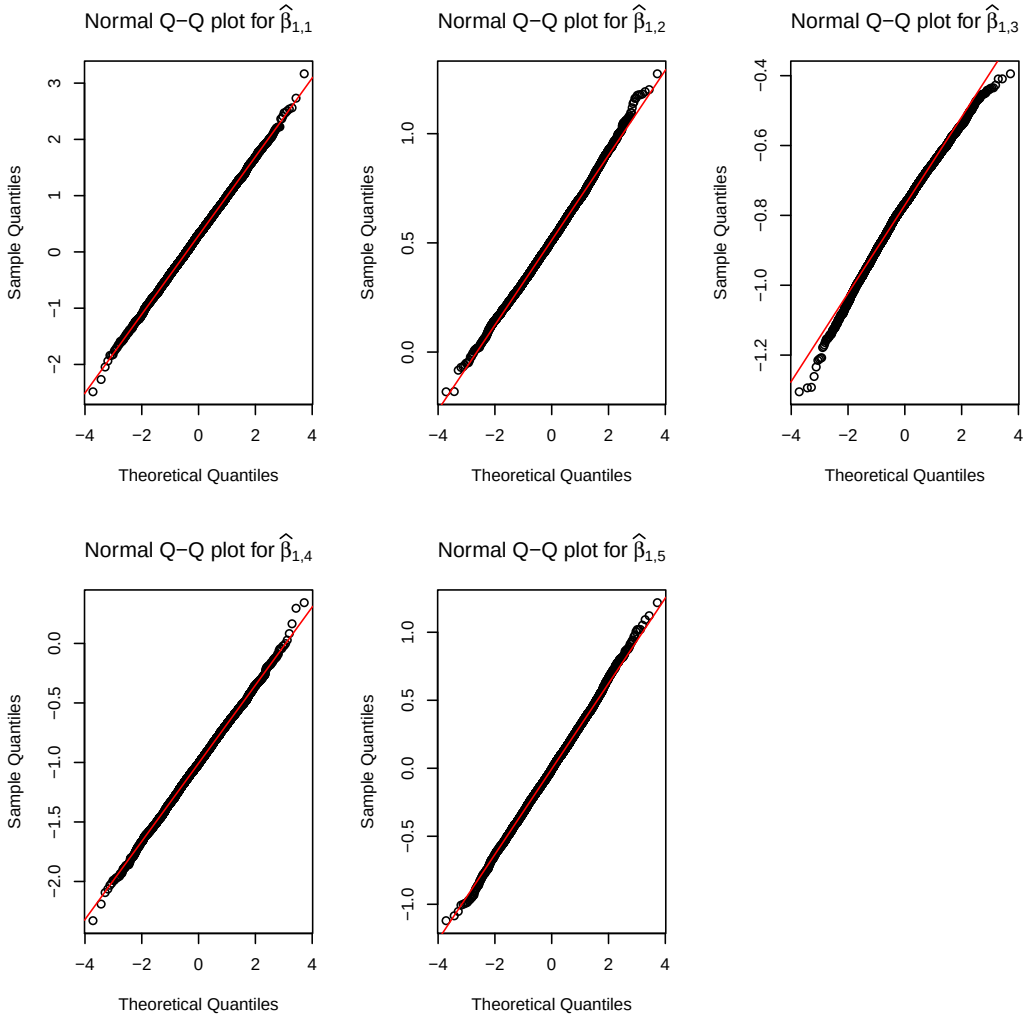

Figure 1: Normal Q-Q plots of $\hat{\beta}_{1,1}, \ldots, \hat{\beta}_{1,5}$, for case (ii) with $n = 300$ and average sample proportion of zero-inflation equal to 0.25.

4. An application in health economics

4.1. Data description and competing models

In this section, we apply the proposed ZIM model to health-care utilization data obtained from the National Medical Expenditure Survey conducted in 1987-1988. This data set was first described by Deb and Trivedi (1997). We consider jointly three health-care utilization measures: the number *ofnd* of consultations with a non-doctor in an office setting, the number *opnd* of consultations with a non-doctor in an outpatient setting and the number *ofd* of consultations with a doctor in an office setting.

The sample contains 3224 individuals with at least two consultations of all types among *ofnd*, *opnd* and *ofd*. Frequencies of individuals with zero occuring simultaneously in (*ofnd* and *opnd*), (*ofnd* and *ofd*) and (*opnd* and *ofd*) are 51.7%, 0.24% and 1% respectively. The high frequency of zeros in (*ofnd*, *opnd*) and low frequencies of zero counts in the other two pairs of health services suggest that there may exist permanent non-users of the combination (*ofnd* and *opnd*). That is, there may exist an excess of observations of the

type $(0, 0, m_i)$, where m_i denotes the total number of consultations for individual i. Hence we propose to use ZIM model to investigate the determinants of health-care utilization in this data set.

Several covariates were recorded on each individual. They include: i) socio-economic variables: gender (1 for female, 0 for male), age (in years, divided by 10), marital status (1 if married, 0 if not married), educational level (number of years of education), income (in ten-thousands of dollars), ii) various measures of health status: number of chronic conditions (cancer, arthritis, diabete...) and a variable indicating self-perceived health level (poor, average, excellent) and iii) a binary variable indicating whether individual is covered by medicaid or not (medicaid is a US health insurance for individuals with limited income and resources, we code it as 1 if the individual is covered and 0 otherwise). Self-perceived health is re-coded as two dummy variables denoted by "health1" (1 if health is perceived as poor, 0 otherwise) and "health2" (1 health is perceived as average, 0 otherwise).

We fit the following three models: i) a multinomial logistic regression model mult$(m_i, \mathbf{p}_i)$, where $\mathbf{p}_i = (p_{1i}, p_{2i}, p_{3i})$ and the p_{ji} are specified as in (2.3), ii) the ZIM model with fixed probability π of $(0, 0, m_i)$-inflation (denoted by ZIM^a thereafter) and iii) the ZIM model with covariate-dependent probability π_i of $(0, 0, m_i)$-inflation (denoted by ZIM^b), where π_i is as in (2.6). Selection of regressors for inclusion in π_i requires some care. Indeed, it was previously observed in various other zero-inflated models that including all available regressors in both count and zero-inflation probabilities can yield lack of identification of model parameters. See for example Diop *et al.* (2011) and Staub and Winkelmann (2013), who suggest to solve this issue by letting at least one of the covariates included in the count model to be excluded from the zero-inflation model (or the converse). Such condition is not required in the ZIM model. However, in order to avoid numerical problems, we propose a two-stage procedure for covariate selection in the zero-inflation and count models. In a first stage, we fit a standard logistic regression model with all available covariates to the binary indicator $\delta_{(0,0,m_i)}$. The resulting model is not a model for zero-inflation since some of the $(0, 0, m_i)$ may arise from the multinomial model mult$(m_i, \mathbf{p}_i)$. However, we expect that this rough procedure will still select a relevant subset of covariates, that will be used in a second stage in the logistic sub-model (2.6) for π_i. Using this procedure and Wald testing, we identify five significant predictors: age, gender, educational level, number of chronic conditions and medicaid status, that are included in π_i.

4.2. Results

Results for the three competing models (standard multinomial, ZIM^a and ZIM^b) are displayed in Table 7. We report estimate, standard error and significance level (as: not significant, significant or very significant) of Wald test for each parameter. For purpose of comparison, we also report log-likelihood and AIC values for the three models. ZIM^b appears as the best model in terms of both likelihood and AIC (in an unreported analysis, we also fitted ZIM^b with various other subsets of covariates in π_i. The smallest AIC is achieved for the subset selected by our two-stage procedure).

Among 1667 non-users of both *ofnd* and *opnd*, 41.5% are identified as permanent non-users by ZIM^a. Gender, educational level and medicaid status are identified by ZIM^b as the most influencing factors for being a permanent non-user, with medicaid recipients being more likely to be permanent non-users. The three models identify the same subset of influent factors for *opnd* utilization, with similar parameter estimates except for medicaid status: ZIM^a and ZIM^b suggest that probability of using *opnd* is less sensitive to medicaid status than suggested by standard multinomial regression. Moreover, for ZIM^a and ZIM^b, medicaid status does not affect *ofnd* utilization. These findings are coherent with the fact that part of the decision of (not) using *ofnd* and *opnd* by medicaid recipients was captured in the model for π_i. All this suggests that medicaid recipients tend to favor doctor visits in an office setting over non-doctor visits in either office or outpatient settings. This confirms previous findings that patients with medicaid insurance coverage have less non-doctor health professional visits. This feature is also captured by the standard multinomial regression model but ZIM model additionally confirms that medicaid recipients are more likely to decide to never use *ofnd* and *opnd* services. From ZIM^b, educational level is an important determinant of the decision of being a permanent non-user of both *ofnd* and *opnd*. But once an individual has chosen to use eventually these health-care services (with a probability that increases with level of education), ZIM^b suggests that schooling does not tend to favor a specific kind of health-care service. Income does not affect utilization of medical

care. This is consistent with previous findings (*e.g.*, Deb and Trivedi, 1997) and is explained in the literature by the fact that income may affect intensity and quality of care rather than visits number. Marital status has a strong effect on *ofnd* and *opnd* utilization, with similar magnitude but opposite sign. Married patients are more likely to visit a doctor in an office setting, which may be due to couples having more financial resources than single individuals. Deb and Trivedi (1997) report that an increase in the number of chronic conditions increases utilization of each form of medical care. We find here that chronic condition does not affect *opnd* and affects negatively *ofnd*. Thus, *ofd* utilization increases with the number of chronic conditions. Contradiction with conclusions by Deb and Trivedi (1997) is only apparent. By considering simultaneously *ofnd*, *opnd* and *ofd*, we are able to rank the various forms of medical care by order of utilization. Our observation reflects the fact that as the number of chronic conditions increases, doctor visits are preferred to non-doctor visits, which seems natural.

5. Conclusion

In this paper, we introduce a model for multivariate count data with excess zeros when zero-inflation affects jointly several component counts. Maximum likelihood estimation is shown to perform well in this model, under a range of scenarios. Moreover, in our analysis of health-care utilization, the proposed model provides plausible explanations and interpretations and gives useful insight into the decision of using or not available health-care services. Several issues now deserve attention, such as derivation of a formal test for zero-inflation in multinomial counts. Generalizing the proposed model to more complex settings (*e.g.*, cluster correlation, longitudinal or hierarchical data) is also desirable and constitutes the topic for our future work.

Appendix A. Proof of identifiability.

Suppose that $l_{[i]}(\psi) = l_{[i]}(\psi^*)$ almost surely. Under C1 and C2, there exists $\epsilon > 0$ such that for every $\mathbf{X}_i$ and $\psi \in \mathbf{K}$, $\epsilon < \mathbb{P}\left(Z_i \neq (0,0,m_i)|\mathbf{X}_i\right) = (1-\pi)\left(1-(h_i(\beta))^{-m_i}\right)$. Therefore, we can find $\omega \in \Omega$, with ω outside the negligible set where $l_{[i]}(\psi) \neq l_{[i]}(\psi^*)$, such that $Z_i(\omega) \neq (0,0,m_i)$. For such ω, $J_i = 1$ and thus, $l_{[i]}(\psi) = l_{[i]}(\psi^*)$ becomes :

$$Z_{1i}(\beta_1 - \beta_1^*)^\top \mathbf{X}_i + Z_{2i}(\beta_2 - \beta_2^*)^\top \mathbf{X}_i = \log\left[\left(\frac{h_i(\beta)}{h_i(\beta^*)}\right)^{m_i} \times \left(\frac{1-\pi^*}{1-\pi}\right)\right]. \tag{5.7}$$

The right-hand side of (5.7) does not depend on $Z_i = (Z_{1i}, Z_{2i}, Z_{3i})$. Therefore, the left-hand side of (5.7) should be constant for two different values of Z_i. Consider for example $Z_i = (z_{1i}, z_{2i}, m_i - z_{1i} - z_{2i})$ and $Z_i = (z_{1i}, z_{2i} - 1, m_i - z_{1i} - z_{2i} + 1)$, with $z_{1i}, z_{2i} \geq 1$ (which is possible since $m_i \geq 2$ by C4). Then we obtain $(\beta_2 - \beta_2^*)^\top \mathbf{X}_i = 0$. A similar argument yields $\beta_1 = \beta_1^*$ and finally, $\pi = \pi^*$, which concludes the proof. □

Appendix B. Proofs of asymptotic results.

An intermediate technical lemma is first proved.

Lemma 5.1. *Let $\phi_n : \mathbb{R}^k \longrightarrow \mathbb{R}^k$ be defined as $\phi_n(\psi) = \psi + (\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\dot{l}_n(\psi)$. Then there exists an open ball $B(\psi_0, r)$ (with $r > 0$) and a constant $0 < c < 1$ such that:*

$$\left\|\phi_n(\psi) - \phi_n(\tilde{\psi})\right\| \leq c\left\|\psi - \tilde{\psi}\right\| \text{ for all } \psi, \tilde{\psi} \in B(\psi_0, r). \tag{5.8}$$

Proof of Lemma 5.1. Property (5.8) holds if we can prove that $\left\|\frac{\partial \phi_n(\psi)}{\partial \psi^\top}\right\| \leq c$ for all $\psi \in B(\psi_0, r)$. We have:

$$\begin{aligned}
\left\|\frac{\partial \phi_n(\psi)}{\partial \psi^\top}\right\| &= \left\|I_k + (\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\ddot{l}_n(\psi)\right\| \\
&= \left\|(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\| \\
&\leq \left\|(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\right\| \left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\| \\
&= \lambda_n^{-1}\left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\|.
\end{aligned}$$

		multinomial model			ZIM^a			ZIM^b		
parameter	variable	est.	s.e.	test	est.	s.e.	test	est.	s.e.	test
$\beta_{1,1}$	intercept	-1.6311	0.2511	VS	-0.8986	0.2887	VS	-0.9331	0.3883	S
$\beta_{1,2}$	health1	-0.8457	0.0919	VS	-0.7275	0.1058	VS	-0.7308	0.1043	VS
$\beta_{1,3}$	health2	-0.3143	0.0681	VS	-0.3089	0.0793	VS	-0.3072	0.0790	VS
$\beta_{1,4}$	chronic	-0.0903	0.0141	VS	-0.1243	0.0161	VS	-0.1270	0.0164	VS
$\beta_{1,5}$	age	-0.0287	0.0301	NS	0.0023	0.0349	NS	0.0214	0.0445	NS
$\beta_{1,6}$	gender	0.3155	0.0407	VS	0.2058	0.0462	VS	0.1839	0.0475	VS
$\beta_{1,7}$	marital status	0.2160	0.0414	VS	0.2028	0.0468	VS	0.2031	0.0473	VS
$\beta_{1,8}$	educational	0.0405	0.0055	VS	0.0152	0.0064	S	0.0071	0.0068	NS
$\beta_{1,9}$	income	-0.0084	0.0061	NS	-0.0098	0.0065	NS	-0.0093	0.0065	NS
$\beta_{1,10}$	medicaid	-0.3406	0.0757	VS	-0.1217	0.0908	NS	-0.0276	0.0893	NS
$\beta_{2,1}$	intercept	1.0023	0.4982	S	1.8090	0.5235	VS	1.7695	0.4370	VS
$\beta_{2,2}$	health1	0.4011	0.1838	S	0.5185	0.1807	VS	0.5102	0.1891	VS
$\beta_{2,3}$	health2	0.4084	0.1611	S	0.4063	0.1567	VS	0.4051	0.1718	S
$\beta_{2,4}$	chronic	-0.0036	0.0232	NS	-0.0339	0.0246	NS	-0.0363	0.0249	NS
$\beta_{2,5}$	age	-0.5980	0.0593	VS	-0.5741	0.0627	VS	-0.5539	0.0519	VS
$\beta_{2,6}$	gender	0.0870	0.0698	NS	-0.0079	0.0729	NS	-0.0301	0.0747	NS
$\beta_{2,7}$	marital status	-0.2317	0.0708	VS	-0.2407	0.0732	VS	-0.2407	0.0754	VS
$\beta_{2,8}$	educational	0.0185	0.0096	NS	-0.0100	0.0105	NS	-0.0180	0.0104	NS
$\beta_{2,9}$	income	0.0113	0.0095	NS	0.0112	0.0095	NS	0.0116	0.0094	NS
$\beta_{2,10}$	medicaid	-0.6667	0.1385	VS	-0.4809	0.1522	VS	-0.3905	0.1605	S
π					0.4150	0.0107				
γ_1	intercept							-0.5814	1.3793	NS
γ_2	chronic							-0.0345	0.0339	NS
γ_3	age							0.1661	0.1706	NS
γ_4	gender							-0.2711	0.0994	VS
γ_5	educational							-0.0763	0.0150	VS
γ_6	medicaid							0.5784	0.1788	VS
log-lik		-15201.11			-14183.48			**-14142.65**		
AIC		30442.22			28408.97			**28337.31**		

Table 7: Health-care data analysis: estimates, log-likelihood and AIC values from multinomial, ZIM^a and ZIM^b models (NS: not significant at the 5% level, S: significant at level between 1% and 5%, VS (very significant): significant at level less than 1%).

Now, let $\mathcal{I}$ denote the set of indices $\{(i,j) \in \{1,2,\ldots,3n\}^2$ such that $\mathbb{D}_{ij}(\psi_0) \neq 0\}$. Then the following holds:

$$\begin{aligned}\left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\| &= \left\|\sum_{i=1}^{3n}\sum_{j=1}^{3n} \mathbb{V}_{\bullet i}\mathbb{V}_{\bullet j}^\top(\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0))\right\| \\ &\leq \sum_{(i,j)\in\mathcal{I}} \left\|\mathbb{V}_{\bullet i}\mathbb{V}_{\bullet j}^\top \mathbb{D}_{ij}(\psi_0)\right\| \left|\frac{\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0)}{\mathbb{D}_{ij}(\psi_0)}\right|.\end{aligned}$$

Under conditions C1 and C2, there exists a constant c_2 ($c_2 > 0$) such that $|\mathbb{D}_{ij}(\psi_0)| > c_2$ for every $(i,j) \in \mathcal{I}$. For example, consider the case where $\mathbb{D}_{ij}(\psi_0)$ coincides with some $\mathbb{D}_{4,\ell\ell}(\psi_0)$, for $\ell \in \{1,\ldots,n\}$. Then

$$|\mathbb{D}_{4,\ell\ell}(\psi)| = \frac{m_\ell e^{\beta_1^\top \mathbf{X}_\ell}(h_\ell(\beta))^{m_\ell - 1}}{\left(\pi[(h_\ell(\beta))^{m_\ell} - 1] + 1\right)^2} > \frac{m_{\mathbf{X}}}{(1 + 2M_{\mathbf{X}})^{2m_\ell}},$$

where $m_{\mathbf{X}} := \min_{\beta,\mathbf{X}} e^{\beta^\top \mathbf{X}}$ and $M_{\mathbf{X}} := \max_{\beta,\mathbf{X}} e^{\beta^\top \mathbf{X}}$. Under C1, C2, C4, there exists a positive constant d_4 such that $\frac{m_{\mathbf{X}}}{(1+2M_{\mathbf{X}})^{2m_\ell}} > d_4$. Using similar arguments, we obtain that $|\mathbb{D}_{i,\ell\ell}(\psi)| > d_i$ for some $d_i, i = 1,\ldots,6$. If $c_2 = \min_{1\leq i\leq 6} d_i$, we obtain $|\mathbb{D}_{ij}(\psi_0)| > c_2$ for every $(i,j) \in \mathcal{I}$. Moreover, $\mathbb{D}_{ij}(\cdot)$ is uniformly continuous on $\mathbf{K}$ thus for every $\epsilon > 0$, there exists a positive r such that for all $\psi \in B(\psi_0, r)$, $|\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0)| < \epsilon$. It follows that

$$\begin{aligned}\left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{V}^\top\right\| &\leq \frac{\epsilon}{c_2}\sum_{(i,j)\in\mathcal{I}} \left\|\mathbb{V}_{\bullet i}\mathbb{V}_{\bullet j}^\top \mathbb{D}_{ij}(\psi_0)\right\| \\ &= \frac{\epsilon}{c_2}\text{trace}\left(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top\right) \\ &\leq \frac{\epsilon}{c_2} k\Lambda_n.\end{aligned}$$

This in turn implies that $\left\|\frac{\partial\phi_n(\psi)}{\partial\psi^\top}\right\| \leq \frac{\epsilon k\Lambda_n}{c_2\lambda_n} < \frac{\epsilon k c_1}{c_2}$. Now, choosing $\epsilon = c\frac{c_2}{kc_1}$ with $0 < c < 1$, we get that $\left\|\frac{\partial\phi_n(\psi)}{\partial\psi^\top}\right\| \leq c$ for all $\psi \in B(\psi_0, r)$, which concludes the proof. □

Proof of Theorem 2.2. Let the function η_n be defined as $\eta_n(\psi) := \psi - \phi_n(\psi) = -(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\dot{l}_n(\psi)$. Then $\eta_n(\psi_0)$ converges almost surely to 0 as $n \to \infty$. To see this, note that

$$\eta_n(\psi_0) = \left(\frac{1}{n}\ddot{l}_n(\psi_0)\right)^{-1} \cdot \left(\frac{1}{n}\dot{l}_n(\psi_0)\right).$$

By C3, $\left(\frac{1}{n}\ddot{l}_n(\psi_0)\right)^{-1}$ converges to some matrix Σ. Moreover,

$$\frac{1}{n}\dot{l}_n(\psi_0) = \frac{1}{n}\mathbb{V}C(\psi_0) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n A_i(\psi_0) \\ \frac{1}{n}\sum_{i=1}^n X_{i1}B_{i,1}(\psi_0) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_{ip}B_{i,1}(\psi_0) \\ \frac{1}{n}\sum_{i=1}^n X_{i1}B_{i,2}(\psi_0) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n X_{ip}B_{i,2}(\psi_0) \end{pmatrix}$$

converges to 0 almost surely as $n \to \infty$. To see this, note that for every $i = 1,\ldots,n$, $\mathbb{E}[A_i(\psi_0)] = \mathbb{E}[\mathbb{E}[A_i(\psi_0)|\mathbf{X}_i]]$, where

$$\mathbb{E}[A_i(\psi_0)|\mathbf{X}_i] = \frac{(h_i(\beta_0))^{m_i} - 1}{\pi_0[(h_i(\beta_0))^{m_i} - 1] + 1}\mathbb{E}\left[1 - J_i|\mathbf{X}_i\right] - \frac{1}{1-\pi_0}\mathbb{E}\left[J_i|\mathbf{X}_i\right].$$

Now,

$$\mathbb{E}\left[J_i|\mathbf{X}_i\right] = \mathbb{P}(Z_i \neq (0,0,m_i)|\mathbf{X}_i) = (1-\pi_0)\left(1 - \frac{1}{(h_i(\beta_0))^{m_i}}\right),$$

thus

$$\begin{aligned}\mathbb{E}[A_i(\psi_0)|\mathbf{X}_i] &= \frac{(h_i(\beta_0))^{m_i}-1}{\pi_0[(h_i(\beta_0))^{m_i}-1]+1}\left[\pi_0 + (1-\pi_0)\frac{1}{(h_i(\beta_0))^{m_i}}\right] - \left(1 - \frac{1}{(h_i(\beta_0))^{m_i}}\right)\\ &= 0.\end{aligned}$$

It follows that $\mathbb{E}[A_i(\psi_0)] = 0$. Next, for every $i = 1, \ldots, n$,

$$\begin{aligned}\mathrm{var}(A_i(\psi_0)) &= \mathbb{E}[\mathrm{var}(A_i(\psi_0)|\mathbf{X}_i)] + \mathrm{var}(\mathbb{E}[A_i(\psi_0)|\mathbf{X}_i])\\ &= \mathbb{E}[\mathrm{var}(A_i(\psi_0)|\mathbf{X}_i)]\\ &\leq c_3 := \mathbb{E}\left[\left(\frac{(h_i(\beta_0))^{m_i}}{(1-\pi_0)\{\pi_0[(h_i(\beta_0))^{m_i}-1]+1\}}\right)^2\right].\end{aligned}$$

Conditions C1, C2, C4 ensure that $c_3 < \infty$ and thus

$$\sum_{i=1}^{\infty}\frac{\mathrm{var}(A_i(\psi_0))}{i^2} \leq c_3\sum_{i=1}^{\infty}\frac{1}{i^2} < \infty.$$

Kolmogorov's strong law of large numbers applies to terms $A_i(\psi_0)$ and implies that

$$\frac{1}{n}\sum_{i=1}^{n}\left\{A_i(\psi_0) - \mathbb{E}\left[A_i(\psi_0)\right]\right\} = \frac{1}{n}\sum_{i=1}^{n}A_i(\psi_0)$$

converges almost surely to 0.

Similarly, for every $i = 1, \ldots, n$ and $j = 1, \ldots, p$, we have $\mathbb{E}[X_{ij}B_{i,1}(\psi_0)] = \mathbb{E}[X_{ij}\mathbb{E}[B_{i,1}(\psi_0)|\mathbf{X}_i]]$, where

$$\mathbb{E}[B_{i,1}(\psi_0)|\mathbf{X}_i] = -(1-\pi_0)\frac{m_i e^{\beta_{1,0}^\top \mathbf{X}_i}}{k_i(\psi_0)}\mathbb{E}\left[1 - J_i|\mathbf{X}_i\right] - \frac{m_i e^{\beta_{1,0}^\top \mathbf{X}_i}}{h_i(\beta_0)}\mathbb{E}\left[J_i|\mathbf{X}_i\right] + \mathbb{E}\left[J_i Z_{1i}|\mathbf{X}_i\right]$$

and

$$\mathbb{E}\left[J_i Z_{1i}|\mathbf{X}_i\right] = (1-\pi_0)m_i\frac{e^{\beta_{1,0}^\top \mathbf{X}_i}}{h_i(\beta_0)}.$$

Straightforward calculations yield $\mathbb{E}[B_{i,1}(\psi_0)|\mathbf{X}_i] = 0$ and thus $\mathbb{E}[X_{ij}B_{i,1}(\psi_0)] = 0$. Moreover,

$$\sum_{i=1}^{\infty}\frac{\mathrm{var}(X_{ij}B_{i,1}(\psi_0))}{i^2} < \infty$$

by similar arguments as above. Therefore, $\frac{1}{n}\sum_{i=1}^{n}X_{ij}B_{i,1}(\psi_0)$ converges almost surely to 0. Similar result holds for $\frac{1}{n}\sum_{i=1}^{n}X_{ij}B_{i,2}(\psi_0)$. Finally, $\frac{1}{n}\dot{l}_n(\psi_0)$ and $\eta_n(\psi_0)$ converge almost surely to 0 as $n \to \infty$.

Now, let ϵ be an arbitrary positive value. Almost sure convergence of $\eta_n(\psi_0)$ implies that for almost every $\omega \in \Omega$, there exists an integer $n(\epsilon, \omega)$ such that for any $n \geq n(\epsilon, \omega)$, $\|\eta_n(\psi_0)\| \leq \epsilon$ or equivalently, $0 \in B(\eta_n(\psi_0), \epsilon)$. In particular, let $\epsilon = (1-c)s$ with $0 < c < 1$ such as in Lemma 5.1. Since ϕ_n satisfies Lipschitz condition (5.8), Lemma 2 of Gouriéroux and Monfort (1981) ensures that there exists an element of $B(\psi_0, s)$ (let denote this element by $\widehat{\psi}_n$) such that $\eta_n(\widehat{\psi}_n) = 0$ that is, $(\mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top)^{-1}\dot{l}_n(\widehat{\psi}_n) = 0$. Condition C3 implies that $\dot{l}_n(\widehat{\psi}_n) = 0$ and that $\widehat{\psi}_n$ is the unique maximizer of l_n.

To summarize, we have shown that for almost every $\omega \in \Omega$ and for every $s > 0$, there exists an integer value $n(s,\omega)$ such that if $n \geq n(s,\omega)$, then the maximum likelihood estimator $\widehat{\psi}_n$ exists and $\|\widehat{\psi}_n - \psi_0\| \leq s$ (that is, $\widehat{\psi}_n$ converges almost surely to ψ_0). $\square$

Proof of Theorem 2.3. A Taylor expansion of the score function yields

$$0 = \dot{l}_n(\widehat{\psi}_n) = \dot{l}_n(\psi_0) + \ddot{l}_n(\widetilde{\psi}_n)(\widehat{\psi}_n - \psi_0),$$

where $\widetilde{\psi}_n$ lies between $\widehat{\psi}_n$ and ψ_0. Thus, $\dot{l}_n(\psi_0) = -\ddot{l}_n(\widetilde{\psi}_n)(\widehat{\psi}_n - \psi_0)$. Letting $\widetilde{\Sigma}_n := -\ddot{l}_n(\widetilde{\psi}_n) = \mathbb{V}\mathbb{D}(\widetilde{\psi}_n)\mathbb{V}^\top$ and $\Sigma_{n,0} := \mathbb{V}\mathbb{D}(\psi_0)\mathbb{V}^\top$, we have:

$$\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0) = \left[\widehat{\Sigma}_n^{\frac{1}{2}}\widetilde{\Sigma}_n^{-\frac{1}{2}}\right]\left[\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}}\right]\Sigma_{n,0}^{-\frac{1}{2}}\left(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0)\right). \tag{5.9}$$

Terms $[\widehat{\Sigma}_n^{\frac{1}{2}}\widetilde{\Sigma}_n^{-\frac{1}{2}}]$ and $[\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}}]$ in (5.9) converge almost surely to I_k. To see this, we show for example that $\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\| \xrightarrow{a.s} 0$ as $n \to \infty$. First, note that

$$\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\right\| \leq \Lambda_n^{\frac{1}{2}}\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\right\|\left\|\Lambda_n^{-\frac{1}{2}}\left(\Sigma_{n,0}^{\frac{1}{2}} - \widetilde{\Sigma}_n^{\frac{1}{2}}\right)\right\|, \tag{5.10}$$

and

$$\Lambda_n^{-1}\left\|\Sigma_{n,0} - \widetilde{\Sigma}_n\right\| = \Lambda_n^{-1}\left\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{V}^\top\right\|.$$

By Theorem 2.2, $\widetilde{\psi}_n$ converges almost surely to ψ_0. Let $\omega \in \Omega$ be outside the negligible set where this convergence does not hold. By the same arguments as in proof of Lemma 5.1, for every $\epsilon > 0$, there exists $n(\epsilon,\omega) \in \mathbb{N}$ such that if $n \geq n(\epsilon,\omega)$, then $\Lambda_n^{-1}\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{V}^\top\| \leq \epsilon$. Thus $\Lambda_n^{-1}\|\mathbb{V}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{V}^\top\|$ converges almost surely to 0. By continuity of the map $A \mapsto A^{\frac{1}{2}}$, $\|\Lambda_n^{-\frac{1}{2}}(\Sigma_{n,0}^{\frac{1}{2}} - \widetilde{\Sigma}_n^{\frac{1}{2}})\|$ converges almost surely to 0. Moreover, for n sufficiently large, there exists $0 < c_4 < \infty$ such that almost surely, $\Lambda_n^{\frac{1}{2}}\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\| \leq c_4\Lambda_n^{\frac{1}{2}}/\lambda_n^{\frac{1}{2}} < c_4 c_1^{\frac{1}{2}}$ (by condition C3). Thus $\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\|$ converges almost surely to 0. Almost sure convergence of $\|\widehat{\Sigma}_n^{\frac{1}{2}}\widetilde{\Sigma}_n^{-\frac{1}{2}} - I_k\|$ to 0 follows by similar arguments.

It remains us to show that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0))$ converges in distribution to the Gaussian vector $\mathcal{N}(0, I_k)$. Note that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0)) = \Sigma_{n,0}^{-\frac{1}{2}}\sum_{j=1}^{3n}\mathbb{V}_{\bullet j}C_j(\psi_0)$. Thus, by Eicker (1966), this convergence holds if we can check that the following conditions are fulfilled: 1) $\max_{1\leq j\leq 3n}\mathbb{V}_{\bullet j}^\top(\mathbb{V}\mathbb{V}^\top)^{-1}\mathbb{V}_{\bullet j} \to 0$ as $n \to \infty$, 2) $\sup_{1\leq j\leq 3n}\mathbb{E}[C_j(\psi_0)^2 1_{\{|C_j(\psi_0)|>c\}}] \to 0$ as $c \to \infty$, 3) $\inf_{1\leq j\leq 3n}\mathbb{E}[C_j(\psi_0)^2] > 0$. Condition 1) follows by noting that

$$0 < \max_{1\leq j\leq 3n}\mathbb{V}_{\bullet j}^\top(\mathbb{V}\mathbb{V}^\top)^{-1}\mathbb{V}_{\bullet j} \leq \max_{1\leq j\leq 3n}\|\mathbb{V}_{\bullet j}\|^2\|(\mathbb{V}\mathbb{V}^\top)^{-1}\| = \max_{1\leq j\leq 3n}\|\mathbb{V}_{\bullet j}\|^2/\tilde{\lambda}_n$$

and that $\|\mathbb{V}_{\bullet j}\|$ is bounded, by C1. Moreover, $1/\tilde{\lambda}_n$ tends to 0 as $n \to \infty$ by C3. Condition 2) follows by noting that the $C_j(\psi_0)$, $j = 1,\ldots,3n$ are bounded under C1, C2, C4. Finally, we note that $\mathbb{E}[C_j(\psi_0)^2] = \mathrm{var}(C_j(\psi_0))$ since $\mathbb{E}[C_j(\psi_0)] = 0$, $j = 1,\ldots,3n$. If $j \in \{1,\ldots,n\}$, $C_j(\psi_0) = A_j(\psi_0)$. Then $\mathrm{var}(C_j(\psi_0)) = \mathrm{var}(A_j(\psi_0)) = \mathbb{E}[\mathrm{var}(A_j(\psi_0)|\mathbf{X}_j)]$. Now,

$$\begin{aligned}
\mathrm{var}(A_j(\psi_0)|\mathbf{X}_j) &= \left(\frac{(h_j(\beta_0))^{m_j}}{(1-\pi_0)[\pi_0((h_j(\beta_0))^{m_j}-1)+1]}\right)^2 \mathrm{var}(J_j|\mathbf{X}_j) \\
&= \left(\frac{(h_j(\beta_0))^{m_j}}{(1-\pi_0)[\pi_0((h_j(\beta_0))^{m_j}-1)+1]}\right)^2 \mathbb{P}(Z_j \neq (0,0,m_j)|\mathbf{X}_j)(1-\mathbb{P}(Z_j \neq (0,0,m_j)|\mathbf{X}_j)) \\
&= \left(\frac{(h_j(\beta_0))^{m_j}}{(1-\pi_0)[\pi_0((h_j(\beta_0))^{m_j}-1)+1]}\right)^2 \left((1-\pi_0)(1-\frac{1}{(h_j(\beta_0))^{m_j}})\right)\left(\pi_0 + (1-\pi_0)\frac{1}{(h_j(\beta_0))^{m_j}}\right),
\end{aligned}$$

and thus, $\text{var}(A_j(\psi_0)|\mathbf{X}_j) > 0$ for every $j = 1, \ldots, n$ by C1, C2, C4. It follows that $\text{var}(C_j(\psi_0)) > 0$ for every $j = 1, \ldots, n$. By similar arguments, $\text{var}(C_j(\psi_0)) > 0$ for every $j = 1, \ldots, 3n$ and condition 3) is satisfied.

To summarize, we have proved that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0))$ converges in distribution to $\mathcal{N}(0, I_k)$. This result combined with Slutsky's theorem and equation (5.9) implies that $\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0)$ converges in distribution to $\mathcal{N}(0, I_k)$. □

Acknowledgements

Authors acknowledge financial support from the "Service de Coopération et d'Action Culturelle" of the French Embassy in Senegal and logistical support from Campus France (French national agency for the promotion of higher education, international student services, and international mobility). Authors also acknowledge grants from CEA-MITIC, an African Center of Excellence in Mathematics, Informatics and ICT implemented by Gaston Berger University (Senegal).

References

References

Bagozzi, B. E., 2015. The baseline-inflated multinomial logit model for international relations research. Conflict Management and Peace Science doi 10.1177/0738894215570422 (to appear).

Deb, P., Trivedi, P. K., 1997. Demand for medical care by the elderly: a finite mixture approach. Journal of Applied Econometrics 12(3), 313-336.

Dietz, E., Böhning, D., 2000. On estimation of the Poisson parameter in zero-modified Poisson models. Computational Statistics & Data Analysis 34(4), 441-459.

Diop, A., Diop, A., Dupuy, J.-F., 2011. Maximum likelihood estimation in the logistic regression model with a cure fraction. Electronic Journal of Statistics 5, 460-483.

Diop, A., Diop, A., Dupuy, J.-F., 2016. Simulation-based inference in a zero-inflated Bernoulli regression model. Communications in Statistics - Simulation and Computation 45(10), 3597-3614.

Eicker, F., 1966. A multivariate central limit theorem for random linear vector forms. The Annals of Mathematical Statistics 37(6), 1825-1828.

Garay, A. M., Hashimoto, E. M., Ortega, E. M. M., Lachos, V. H., 2011. On estimation and influence diagnostics for zero-inflated negative binomial regression models. Computational Statistics & Data Analysis 55(3), 1304-1318.

Gouriéroux, C., Monfort, A., 1981. Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. Journal of Econometrics 102, 17, 83-97.

Gurmu, S., Elder, J., 2000. Generalized bivariate count data regression models. Economics Letters 68, 31-36.

Hall, D. B., 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics 56(4), 1030-1039.

Henningsen, A., Toomet, O., 2011. maxLik: A package for maximum likelihood estimation in R. Computational Statistics 26(3), 443-458.

Kelley, M. E., Anderson, S. J., 2008. Zero inflation in ordinal data: Incorporating susceptibility to response through the use of a mixture model. Statistics in Medicine 27, 3674-3688.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1-14.

Li, C.-S., 2011. A lack-of-fit test for parametric zero-inflated Poisson models. Journal of Statistical Computation and Simulation 81(9), 1081-1098.

Lim, H. K., Li, W. K., Yu, P. L. H., 2006. Zero-inflated Poisson regression mixture model. Computational Statistics & Data Analysis 71, 151-158.

Moghimbeigi, A., Eshraghian, M. R., Mohammad, K., McArdle, B., 2008. Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. Journal of Applied Statistics 35(9), 1193-1202.

Monod, A., 2014. Random effects modeling and the zero-inflated Poisson distribution. Communications in Statistics. Theory and Methods 43 (4), 664-680.

Mwalili, S. M., Lesaffre, E., Declerck, D., 2008. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. Statistical Methods in Medical Research 17(2), 123-139.

Ridout, M., Hinde, J., Demetrio, C. G. B., 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics 57(1), 219-223.

Sari, N., 2009. Physical inactivity and its impact on healthcare utilization. Health Economics 18(8), 885-901.

Sarma, S., Simpson, W., 2006. A microeconometric analysis of Canadian health care utilization. Health Economics 15(3), 219-239.

Staub, K. E., Winkelmann, R., 2013. Consistent estimation of zero-inflated count models. Health Economics 22(6), 673-686.

Vieira, A. M. C., Hinde, J. P., Demetrio, C. G. B., 2000. Zero-inflated proportion data models applied to a biological control assay. Journal of Applied Statistics 27(3), 373-389.

Wang, P., 2003. A bivariate zero-inflated negative binomial regression model for count data with excess zeros. Economics Letters 78, 373-378.

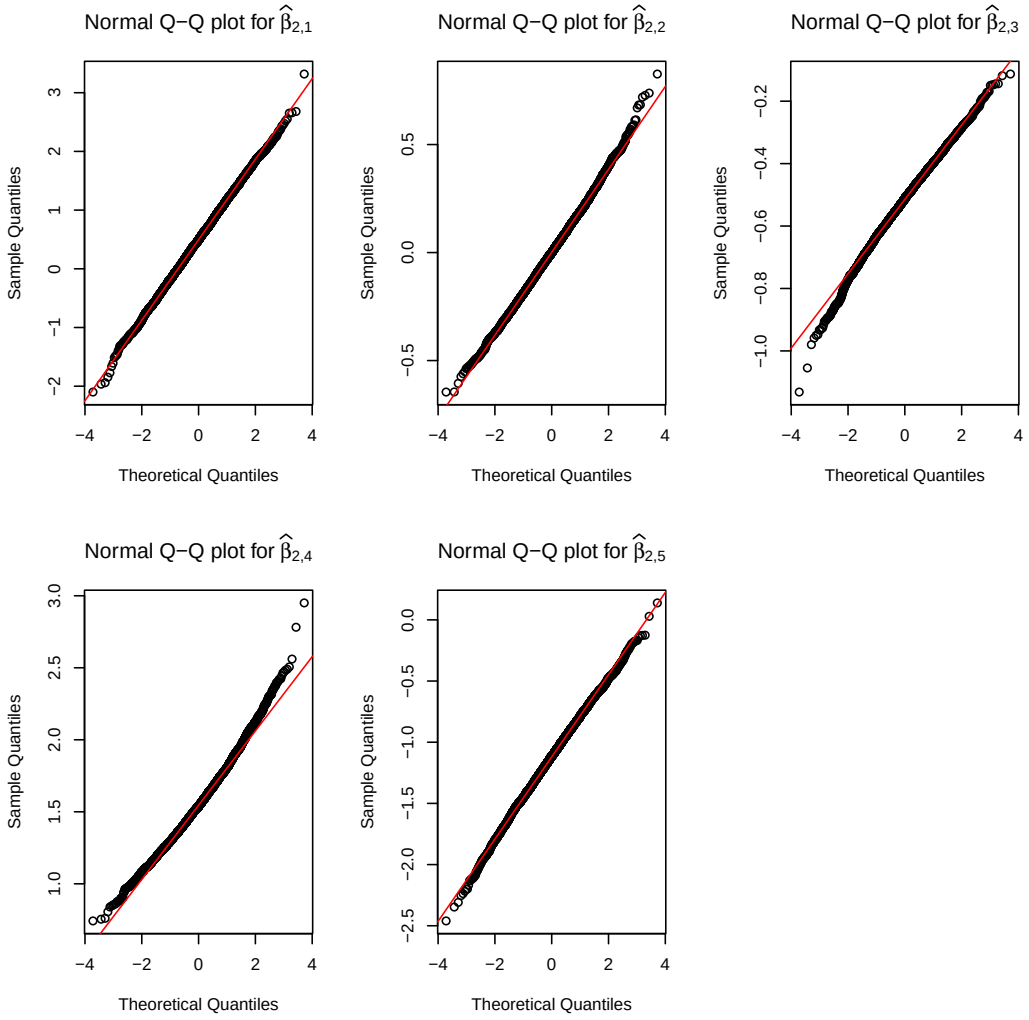

Figure 2: Normal Q-Q plots of $\hat{\beta}_{2,1}, \ldots, \hat{\beta}_{2,5}$, for case (ii) with $n = 300$ and average sample proportion of zero-inflation equal to 0.25.

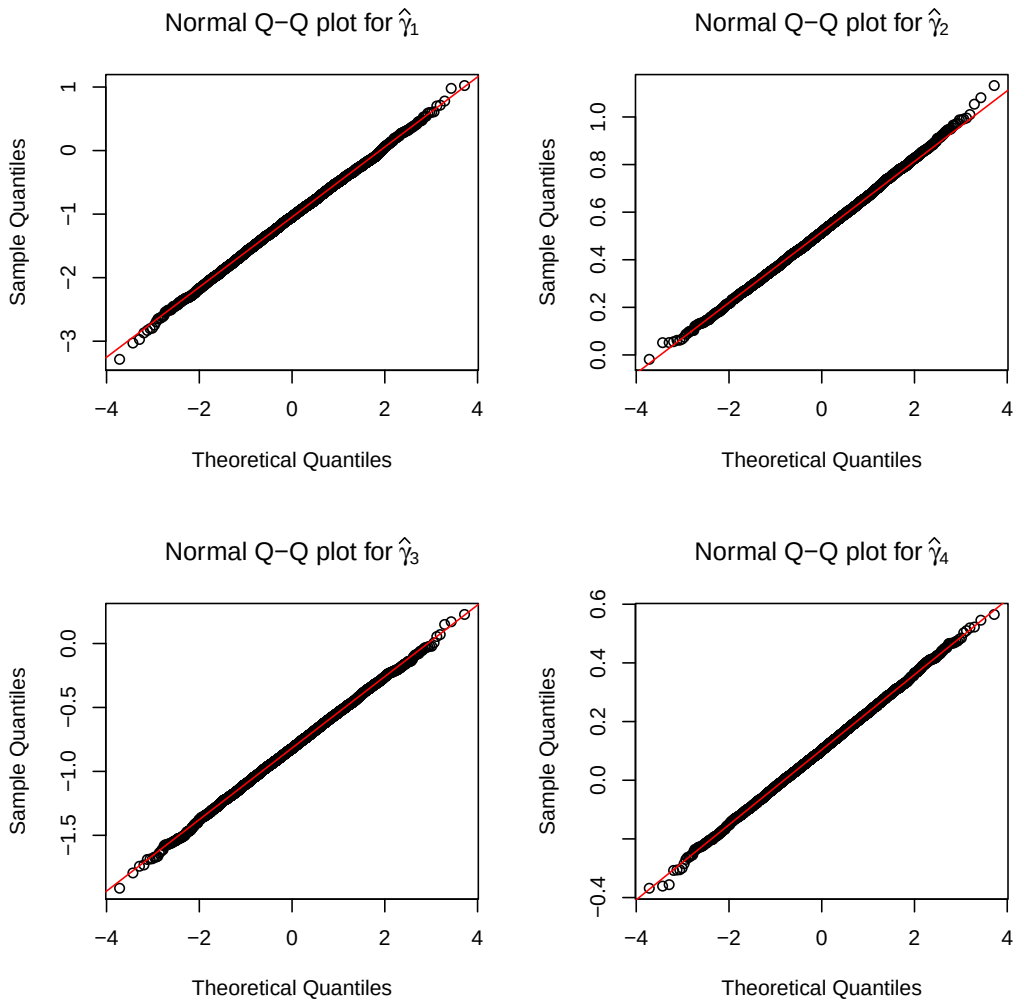

Figure 3: Normal Q-Q plots of $\hat{\gamma}_1, \ldots, \hat{\gamma}_4$, for case (ii) with $n = 300$ and average sample proportion of zero-inflation equal to 0.25.